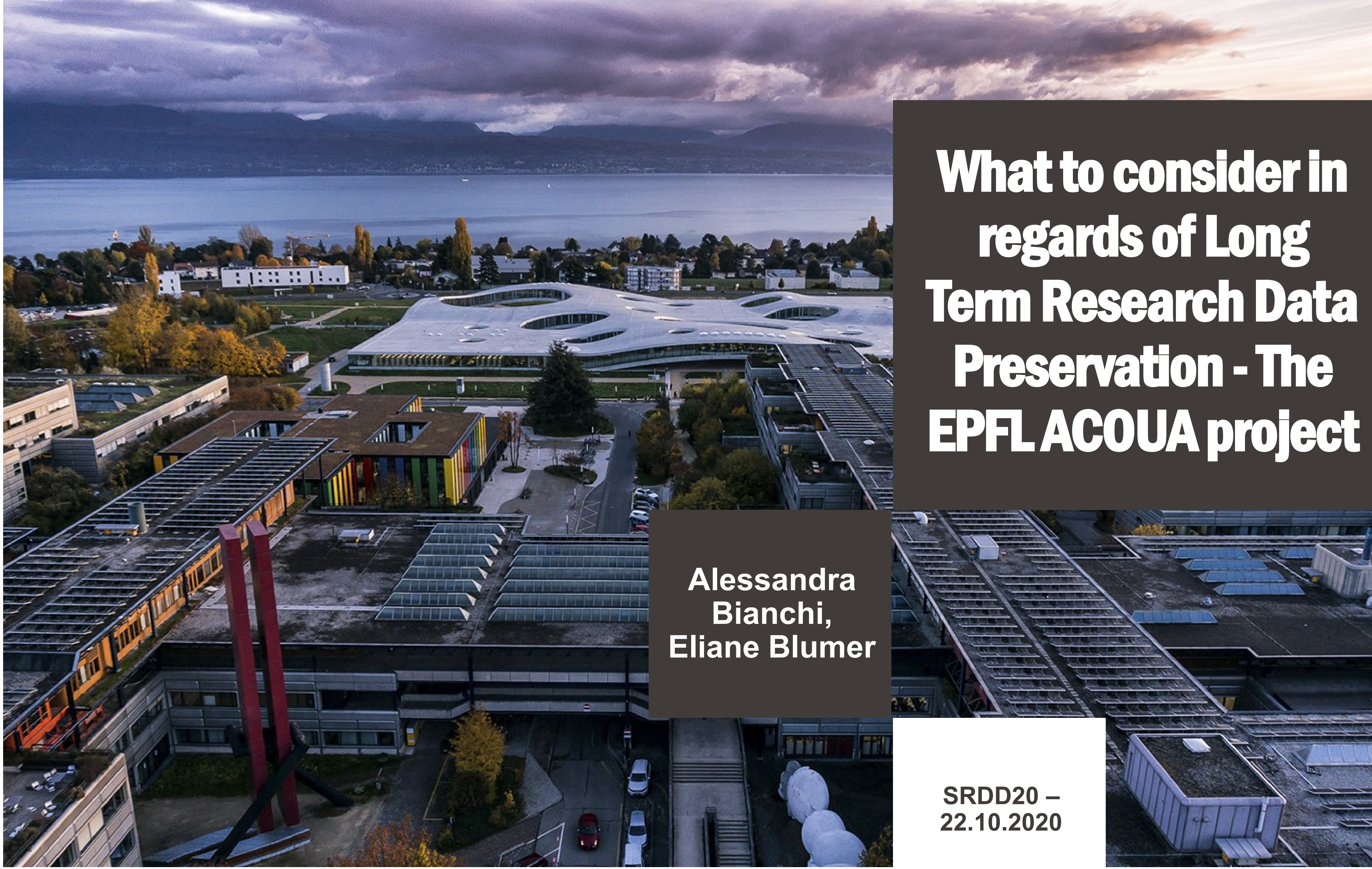


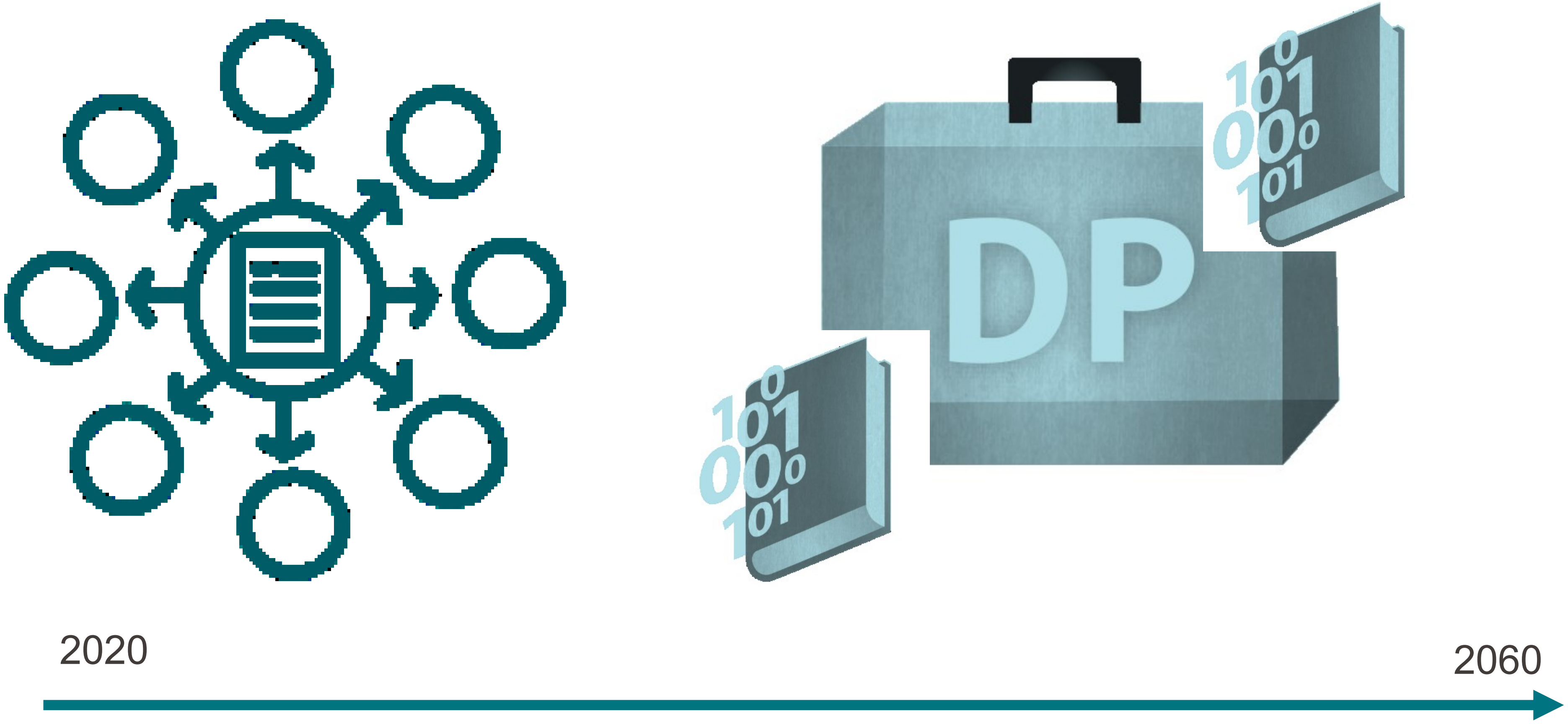
What to consider in regards of Long Term Research Data Preservation - The EPFL ACOUA project

Alessandra Bianchi, Eliane Blumer

SRDD20 – 22.10.2020

École polytechnique fédérale de Lausanne

Context: Data Publication vs. Data Preservation



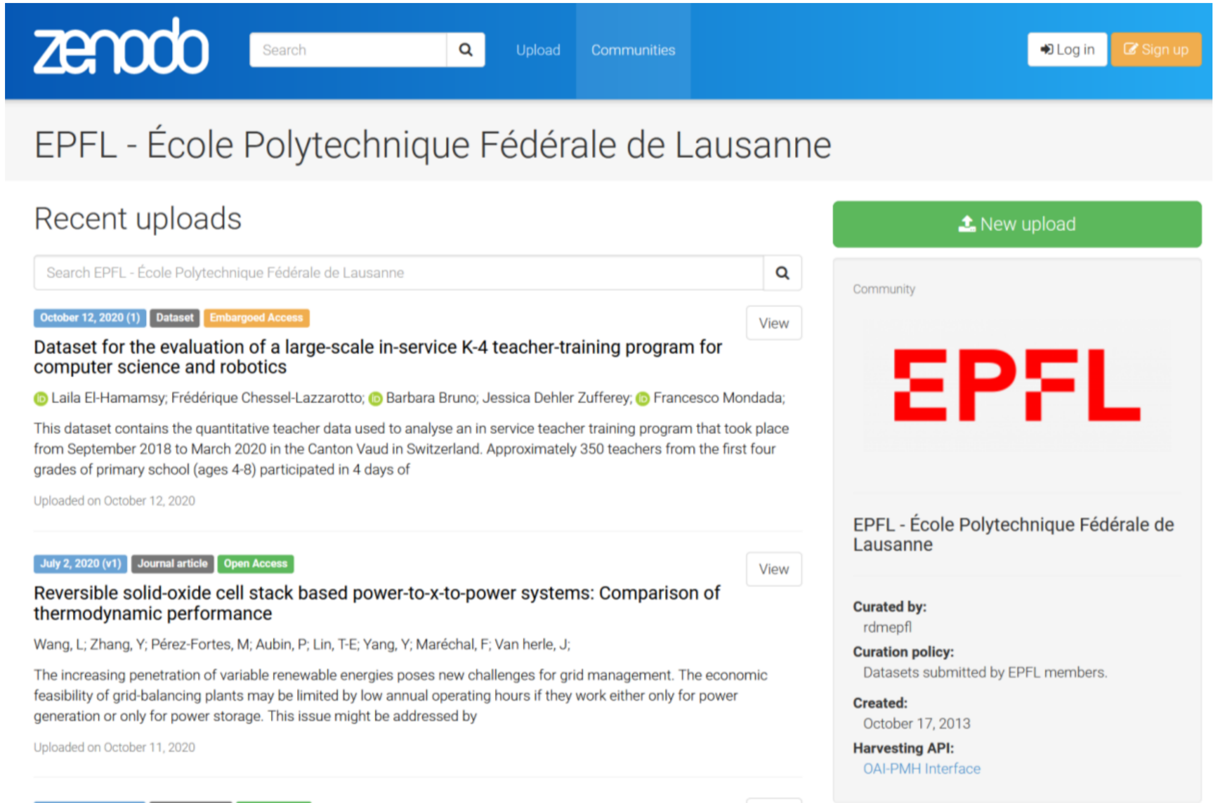
Context: Paradigm Change

- Data is considered a scientific outcome worth to be preserved beyond 10 years
  - Researcher: accessibility and reusability needs to be guaranteed beyond a researcher’s career
  - Institution: history of important outcomes needs to be preserved
  - Funder/Reviewer: 10 years necessity of storage, but what beyond? Is every output considered as «removable»?
  - Community: reusability and transparency on the long run

EPFL ACOUA project history

- Many datasets produced at EPFL are published, but under 404-error threats on the long run
- Several 100 TBs of data existing that need to be preserved
- Institutional requirement for data preservation AND data publication

“And, does the EPFL have an institutional data repository for the storage of data after the project during approx. 10 years?”



EPFL - École Polytechnique Fédérale de Lausanne

Recent uploads

October 12, 2020 (1) Dataset Embargoed Access

Dataset for the evaluation of a large-scale in-service K-4 teacher-training program for computer science and robotics

Laila El-Hamamsy; Frédérique Chessel-Lazzarotto; Barbara Bruno; Jessica Dehler Zufferey; Francesco Mondada;

This dataset contains the quantitative teacher data used to analyse an in service teacher training program that took place from September 2018 to March 2020 in the Canton Vaud in Switzerland. Approximately 350 teachers from the first four grades of primary school (ages 4-8) participated in 4 days of

Uploaded on October 12, 2020

July 2, 2020 (v1) Journal article Open Access

Reversible solid-oxide cell stack based power-to-x-to-power systems: Comparison of thermodynamic performance

Wang, L; Zhang, Y; Pérez-Fortes, M; Aubin, P; Lin, T-E; Yang, Y; Maréchal, F; Van herle, J;

The increasing penetration of variable renewable energies poses new challenges for grid management. The economic feasibility of grid-balancing plants may be limited by low annual operating hours if they work either only for power generation or only for power storage. This issue might be addressed by

Uploaded on October 11, 2020

EPFL - École Polytechnique Fédérale de Lausanne

Curated by:
rdmepfl

Curation policy:
Datasets submitted by EPFL members.

Created:
October 17, 2013

Harvesting API:
OAI-PMH Interface

Main consideration #1: Large volumes

- Large volumes: Do we need to appraise? How are we going to appraise?
  - Currently: data underlying publications
- Problem of price, volume, and speed of transactions

Main consideration #2: variety of formats

- Highly specific and various technical infrastructures are used to produce data

- Proprietary and potentially obsolete formats to be managed

APPROPRIATE	ACCEPTABLE	NOT SUITABLE
.csv - .hdf5	.txt - .html - .tex - .por	
.csv - .tab - .ods - SQL	.xml if appropriate DTD - .xlsx	.xls - .xlsb
.pdf - .txt - .odt - .odm - .tex - .md -.htm - .xml	.pptx - .pdf with embedded forms - .rtf	.doc - .ppt
.m - .R - .py - .iypnb - .rstudio - .rmd – NetCDF	.sdd	.mat - .rdata
.tif - .png - .svg - .jpeg	jpg - .jp2 - .tif - .tiff - .pdf - .gif - .bmp	.indd - .ait - .psd
.flac - .wav - .ogg	.mp3 - .mp4 - .aif	
.mp4 - .mj2 - .avi - .mkv	.ogm - .webm	.wmv - .mov
NetCDF, tabular GIS attribute data, .shp - .shx - .dbf - .prj - .sbx - .sbn - PostGIS - .tif - .tfw - GeoJSON	.mdb - .mif	
.x3d - .x3dv - .x3db - PDF3D .pdf	.dwg - .dxf	
.xml - .json - .rdf		

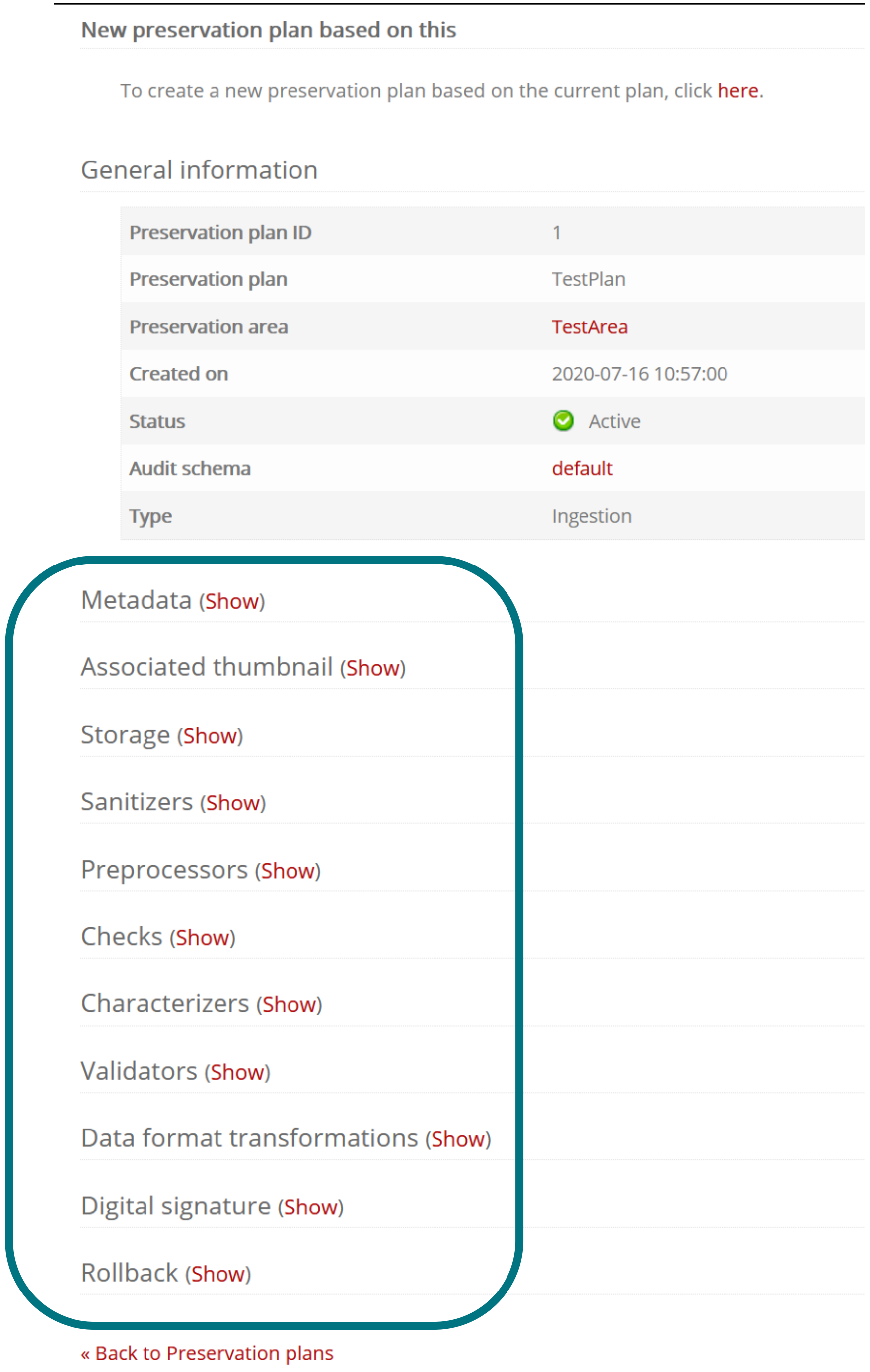
Main consideration #3: Sustainability

- External stakeholders, such as funders’ requirements
- Need for institutional positioning
    - Institutional vs. national policies

Current EPFL situation – ACOUA Project

- Call for tender 2019 for data publication and data preservation
  - Results “data publication”
    - no mature tools for our needs in the context of this call for tender
  - Results “data preservation”
    - Winner is compliant with OAIS-standard, experience with national libraries, co-construction for research data preservation
    - national solutions not mature for our needs in the context of this call for tender

Highly parametrizable preservation tools



New preservation plan based on this

To create a new preservation plan based on the current plan, click here.

General information

Preservation plan ID	1
Preservation plan	TestPlan
Preservation area	TestArea
Created on	2020-07-16 10:57:00
Status	✓ Active
Audit schema	default
Type	Ingestion

Metadata (Show)

Associated thumbnail (Show)

Storage (Show)

Sanitizers (Show)

Preprocessors (Show)

Checks (Show)

Characterizers (Show)

Validators (Show)

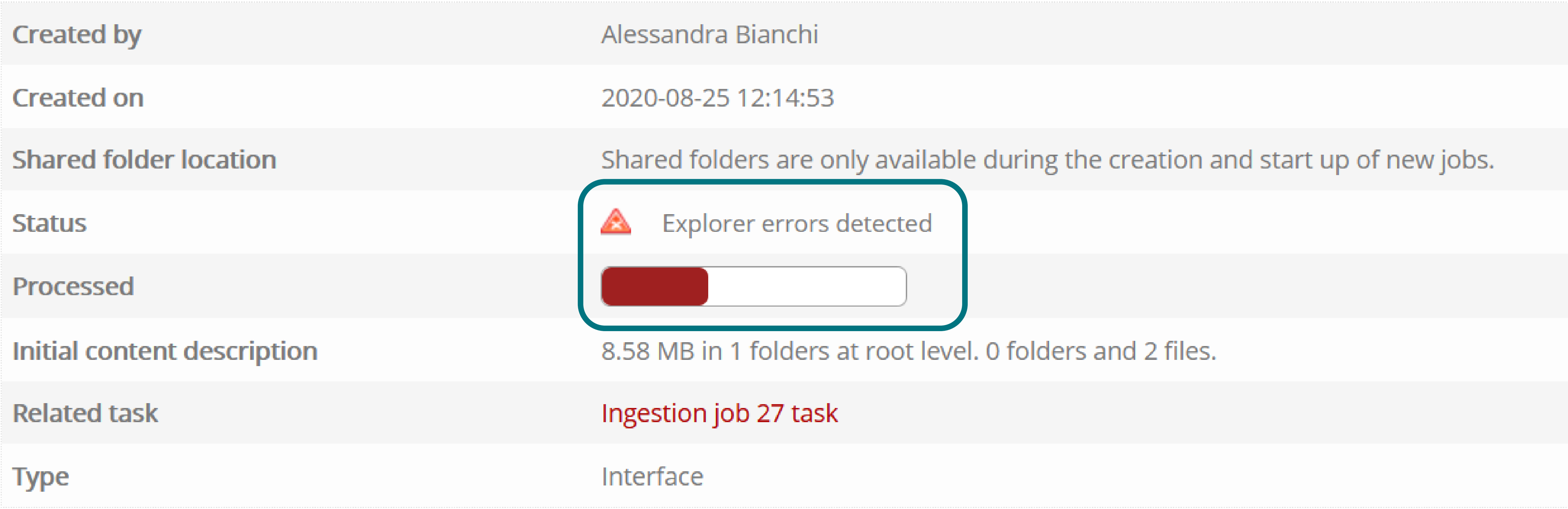
Data format transformations (Show)

Digital signature (Show)

Rollback (Show)

« Back to Preservation plans

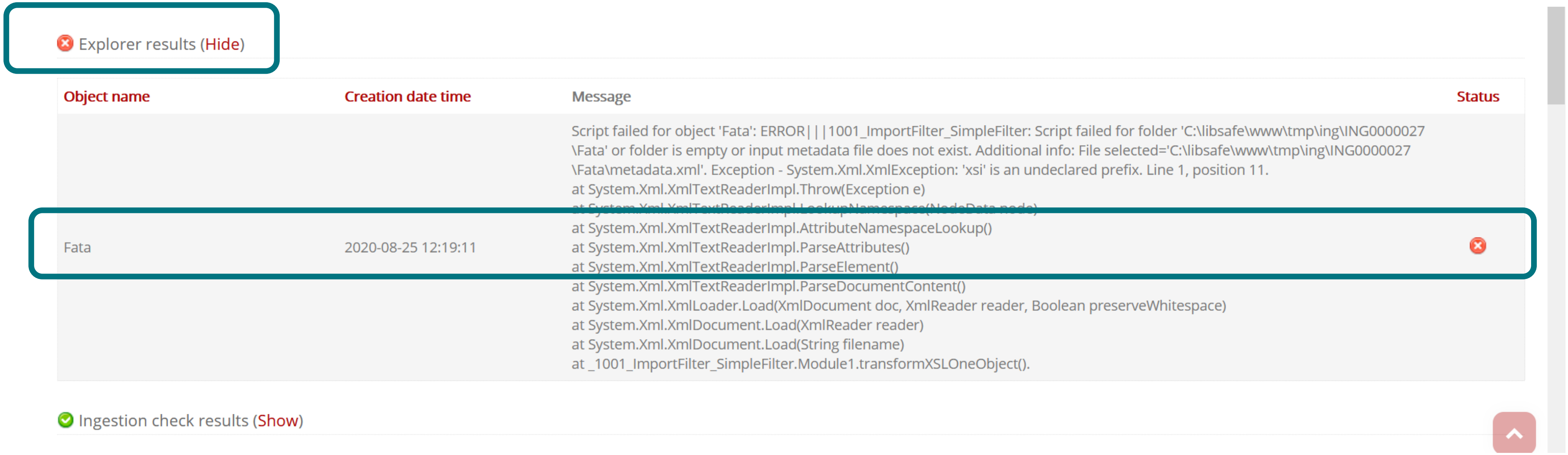
Detailed feedback of failures



Created by	Alessandra Bianchi
Created on	2020-08-25 12:14:53
Shared folder location	Shared folders are only available during the creation and start up of new jobs.
Status	⚠ Explorer errors detected
Processed	
Initial content description	8.58 MB in 1 folders at root level. 0 folders and 2 files.
Related task	Ingestion job 27 task
Type	Interface

Alessandra Bianchi, Eliane Blumer

Detailed feedback of ingest failures

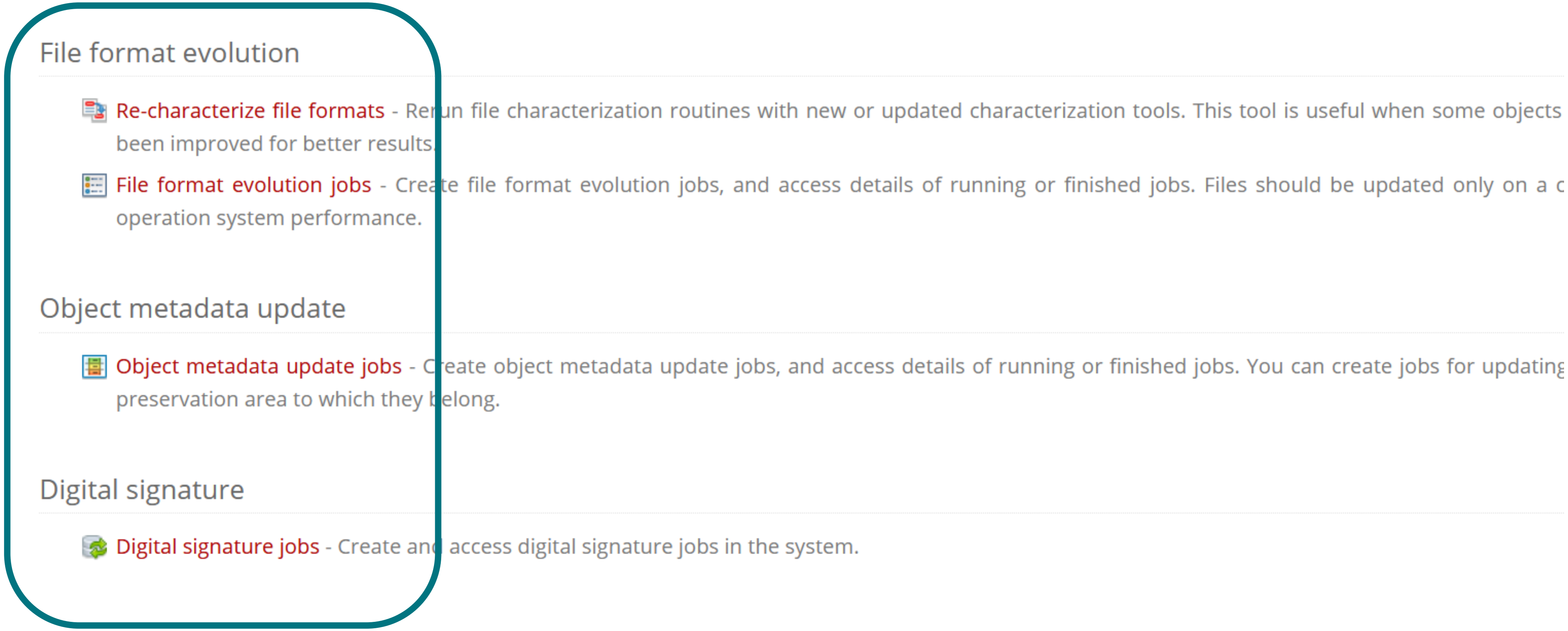


☒ Explorer results (Hide)

Object name	Creation date time	Message	Status
Fata	2020-08-25 12:19:11	Script failed for object 'Fata': ERROR|||1001_ImportFilter_SimpleFilter: Script failed for folder 'C:\libsafe\www\tmp\ing\ING0000027 \Fata' or folder is empty or input metadata file does not exist. Additional info: File selected='C:\libsafe\www\tmp\ing\ING0000027 \Fata\metadata.xml'. Exception - System.Xml.XmlException: 'xsi' is an undeclared prefix. Line 1, position 11. at System.Xml.XmlTextReaderImpl.Throw(Exception e) at System.Xml.XmlTextReaderImpl.LookupNamespace(NodeData node) at System.Xml.XmlTextReaderImpl.AttributeNamespaceLookup() at System.Xml.XmlTextReaderImpl.ParseAttributes() at System.Xml.XmlTextReaderImpl.ParseElement() at System.Xml.XmlTextReaderImpl.ParseDocumentContent() at System.Xml.XmlLoader.Load(XmlDocument doc, XmlReader reader, Boolean preserveWhitespace) at System.Xml.XmlDocument.Load(XmlReader reader) at System.Xml.XmlDocument.Load(String filename) at _1001_ImportFilter_SimpleFilter.Module1.transformXSLOneObject().	☒

✓ Ingestion check results (Show)

Customizable active preservation



File format evolution

Re-characterize file formats - Rerun file characterization routines with new or updated characterization tools. This tool is useful when some objects been improved for better results.

File format evolution jobs - Create file format evolution jobs, and access details of running or finished jobs. Files should be updated only on a c operation system performance.

Object metadata update

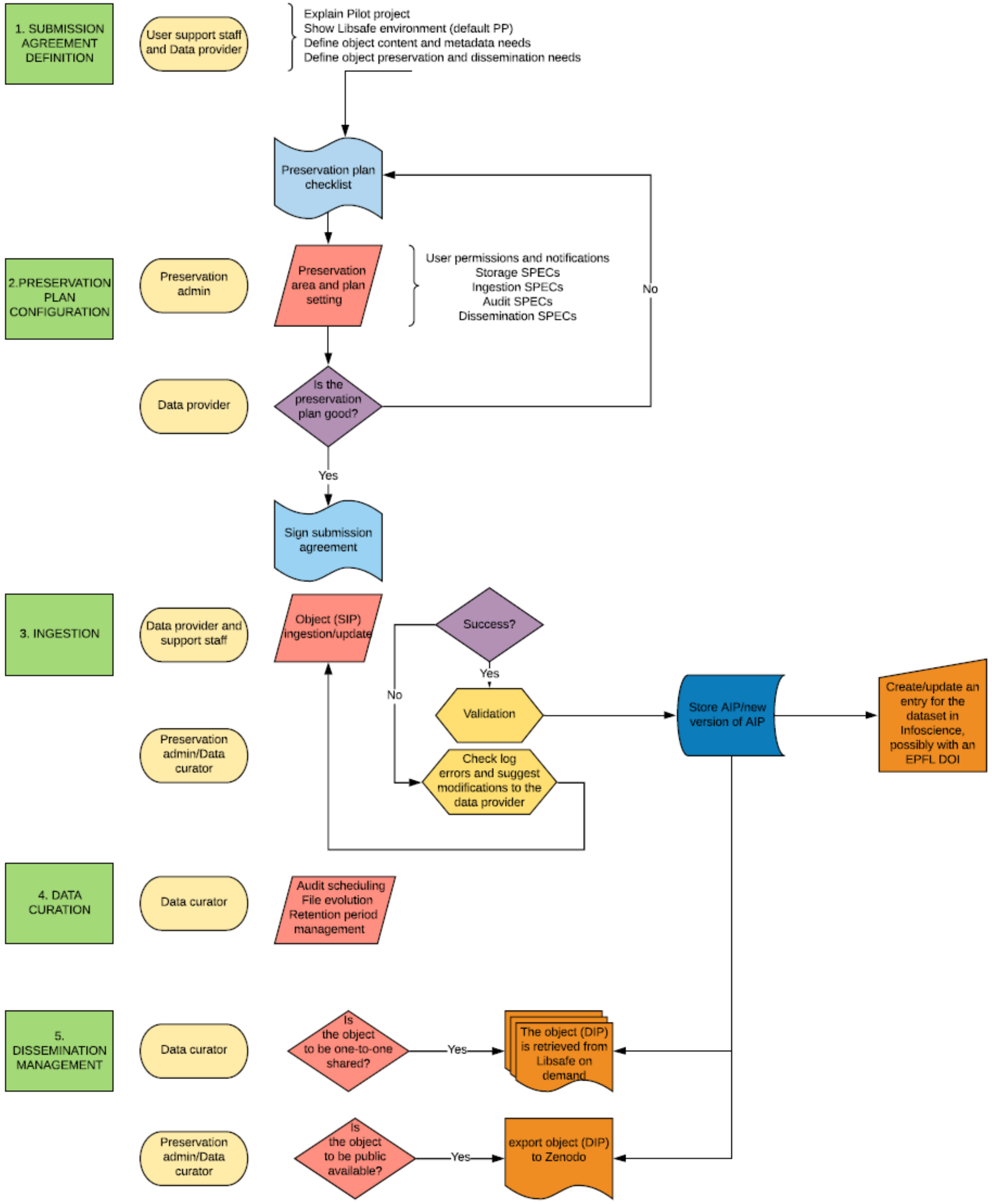
Object metadata update jobs - Create object metadata update jobs, and access details of running or finished jobs. You can create jobs for updating preservation area to which they belong.

Digital signature

Digital signature jobs - Create and access digital signature jobs in the system.

Current EPFL workflow

- Internal preparation and discussions
  - Prepare Submission Agreement
  - Prepare Metadata Application Profile
  - Defining our internal process
  - A lot of project management documentation…
- Pilot with labs going on
- Official switch to PROD: 2021



1. SUBMISSION AGREEMENT DEFINITION

User support staff and Data provider

Explain Pilot project
Show Libsafe environment (default PP)
Define object content and metadata needs
Define object preservation and dissemination needs

Preservation plan checklist

2.PRESERVATION PLAN CONFIGURATION

Preservation admin

Preservation area and plan setting

User permissions and notifications
Storage SPECs
Ingestion SPECs
Audit SPECs
Dissemination SPECs

Data provider

Is the preservation plan good?

Sign submission agreement

3. INGESTION

Data provider and support staff

Object (SIP) ingestion/update

Success?

Validation

Check log errors and suggest modifications to the data provider

Preservation admin/Data curator

Store AIP/new version of AIP

Create/update an entry for the dataset in Infoscience, possibly with an EPFL DOI

4. DATA CURATION

Data curator

Audit scheduling
File evolution
Retention period management

5. DISSEMINATION MANAGEMENT

Data curator

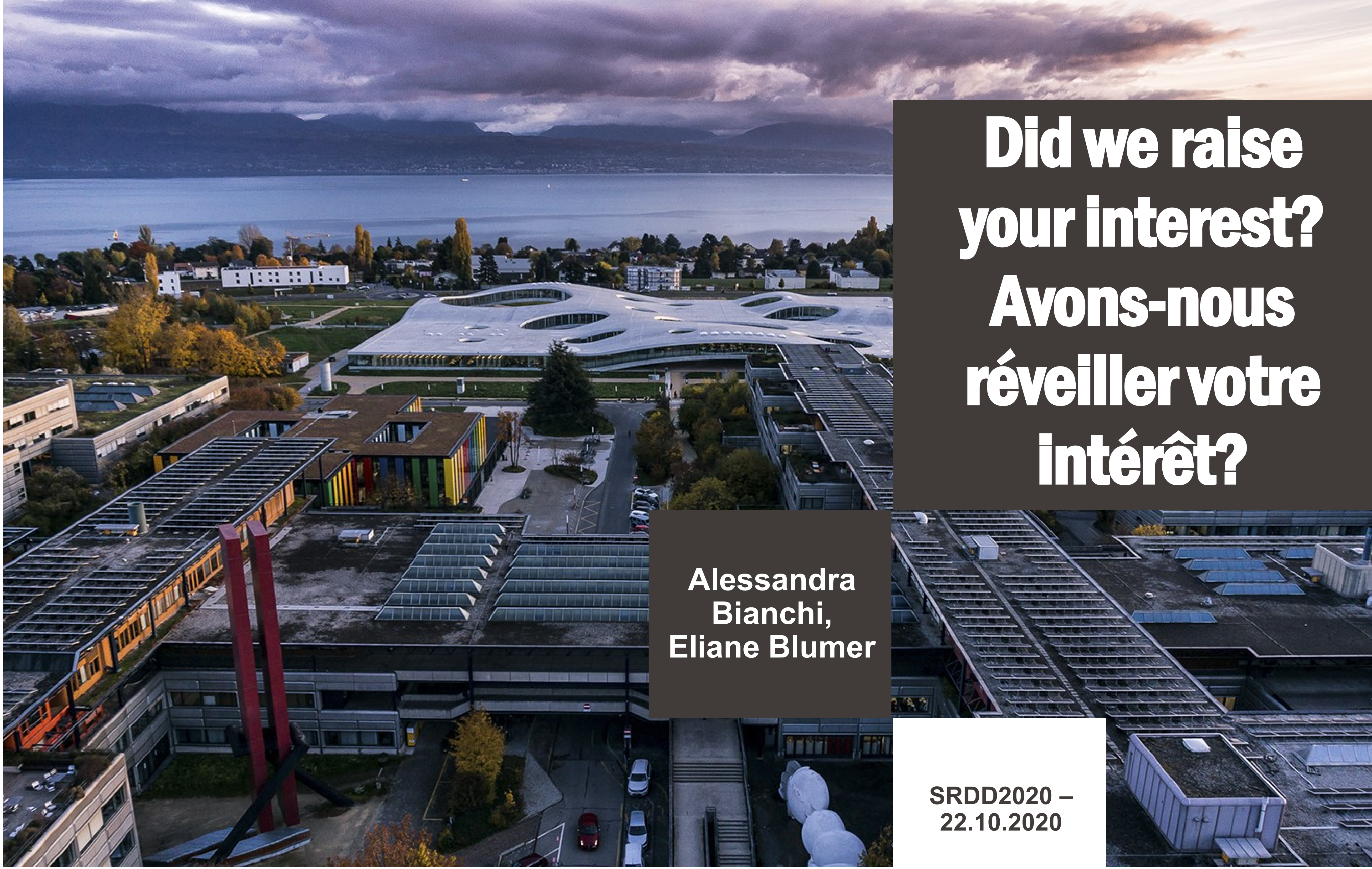
Is the object to be one-to-one shared?

The object (DIP) is retrieved from Libsafe on demand

Preservation admin/Data curator

Is the object to be public available?

export object (DIP) to Zenodo



Did we raise your interest? Avons-nous réveiller votre intérêt?

Alessandra Bianchi, Eliane Blumer

SRDD2020 – 22.10.2020

École polytechnique fédérale de Lausanne